ED 445 000 TM 031 536

AUTHOR	Tay-Lim, Brenda Siok-Hoon; Stone, Clement A.
TITLE	Assessing the Dimensionality of Constructed-Response Tests Using Hierarchical Cluster Analysis: A Monte Carlo Study.
PUB DATE	2000-04-00
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Constructed Response; Factor Analysis; *Item Response Theory; Monte Carlo Methods
IDENTIFIERS	*Dimensionality (Tests); *Hierarchical Cluster Analysis

ABSTRACT

This study explored two methods that are used to assess the dimensionality of item response data. The paper begins with a discussion of the assessment dimensionality and the use of factor-analytic procedures. A number of problems associated with using linear factor analyses to assess dimensionality are also considered. A procedure is presented for hierarchical cluster analysis in combination with a new proximity measure. A simulation was performed to study how well the different cluster methods (group average, centroid, and Ward's cluster method) recovered unidimensional and multidimensional data and whether different cluster methods over- or underestimated the number of dimensions in unidimensional or multidimensional data. In the simulation, only the centroid cluster method recovered the true dimensionality of simulated unidimensional data reasonably well and only in shorter tests. For all other conditions, the three cluster methods consistently overshadowed the true dimensionality of the simulated data. For three-dimensional data, Ward's cluster method was the best performing, and only the group average and Ward's cluster method recovered the multidimensional data well. Implications for practitioners are discussed. (Contains 6 tables and 46 references.) (SLD)

Assessing the dimensionality of constructed-response tests using hierarchical cluster analysis: A Monte Carlo study

By

Brenda Siok-Hoon Tay-Lim
Educational Testing Services

And

Clement A. Stone
University of Pittsburgh

Paper presented at the annual meeting of the American Educational Research Association at New Orleans, April 2000.

Introduction

For a long time, achievement tests were built with dichotomously scored multiple-choice items. Due to the movement towards testing higher order thinking and problem solving skills, large-scale testing programs are devoting more attention to constructed-response items (Bennett, 1993). A constructed-response test is often created based on a high level of dimensionality-based substantive considerations (e.g. content and cognitive process requirements). In addition, different characteristics in a constructed-response test (e.g. number of score levels) may also contribute to construct-irrelevant variance and affect the dimensionality of the test. In constructed-response tests that integrate various levels of performance standards and item characteristics, the dimensionality underlying the test must be evaluated in order to validate any inferences about performance in targeted domains.

Finally, the application of IRT models to constructed-response tests also requires exploring the dimensionality of a set of items. Like most mathematical models, IRT models have a number of assumptions. One assumption is unidimensionality. Therefore assessment of unidimensionality of item response data is essential prior to the application of any unidimensional model if valid inferences are to be drawn from the examinee's standing on the trait of interest (Nandakumar & Yu, 1996).

The purpose of this paper is to explore two different methods that are used to assess dimensionality of item response data. This paper begins with a discussion of the assessment of dimensionality and the use of factor-analytic procedures. A number of problems associated with using linear factor analyses to assess dimensionality are also considered. Hierarchical cluster analysis in combination with a new proximity measure, has been recommended as an alternative procedure for exploring the underlying dimension and structure of constructed-response tests. This procedure is presented, including a discussion of proximity measures or the statistics used to measure the similarity between items; cluster methods or methods used to combine items into clusters or sets; and, stopping criteria or statistics used to determine the optimal number of clusters. Finally, the research questions are presented, followed by methodology, results and analyses, conclusion, implication for practioners, limitations and future research.

Assessing test dimensionality using factor analysis

There are many procedures used to assess the dimensionality of a test, which range from examining measures of internal consistency (McDonald, 1981) to evaluating eigenvalue plots (Lord & Novick, 1968; Reckase, 1979). Factor analysis is a commonly used method to assess dimensionality in a test. Based on past research, there are problems in using factor analysis for examining dimensionality for dichotomous items. First, the phi correlation (special case of Pearson correlation) confounds item difficulty with correlation (Mislevy, 1986). When item difficulties are not uniform, factor analysis often identifies spurious factors. Although the problem associated with phi correlations is alleviated by using tetrachoric correlations, the matrix of sample tetrachoric correlation coefficients is almost never positive definite (Lord & Novick, 1968; Muraki & Engelhard, 1985; Bock, Gibbons, & Muraki, 1988). In addition, the present method of calculating tetrachoric coefficients becomes unstable as the value approaches +1 or –1. When an observed frequency in the 2×2 contingency table for a pair of items is 0, the absolute value of an element in the item correlation matrix becomes 1 and produces the Heywood case or negative covariances (Carroll, 1945). The problem of Heywood cases for tetrachoric correlations may be extended to polychoric correlations.

A second problem with factor analytic methods is that the value of dichotomous responses is bounded, that is, there is only a score of 0 or 1 on the items. This causes the relationship between the item scores and the continuous latent variables (abilities) to be nonlinear (Mislevy, 1986). However, as the number of categories in the item increases, as in the case of polytomous items, the relationship between item scores may become more linear thus alleviating the problem of non-linearity.

Assessing dimensionality using hierarchical cluster analysis

New dimensionality assessment tools have been explored and developed in recent years. One promising tool is hierarchical cluster analysis (HCA). Hierarchical cluster analysis has recently been used as an exploratory tool to investigate the descriptive nature (e.g. content or item type) of items in its identified clusters (Roussos, Stout, & Marden, 1997). It has been applied successfully to standardized tests with dichotomous items (Douglas, Kim, Roussos, Stout, & Zhang, 1995; Roussos & Stout, 1995; Roussos, 1995; Roussos, Stout & Marden, 1997).

I. Proximity measure

The concept of a cluster is closely linked to the concept of proximity between objects and groups of objects. A common measure of similarity is based on the Pearson correlation coefficient. These types of proximity measures are sometimes called correlation-type measures. Dissimilarity measures are sometime referred as distance-type measures. A commonly used measure of dissimilarity is the Euclidean distance, which is the sum of distance differences between two points.

In test-response data, the proximity measures chosen or developed have to be sensitive to differences in dimensionality between items. However, the classical proximity measures tend to cluster items on the basis of difficulty level rather than dimensionality. The common correlation-type and distance-type measures are therefore not sufficient to use as proximity measures for the purpose of present study.

To address this problem, Roussos (1995, 1997) developed a new proximity measure based on research by Cronbach and Gleser (1953), Hambleton & Rovinelli (1986), McDonald (1982), Roznowski, Tucker, & Humphreys (1991), and Roussos, Stout, & Marden (1997). In this approach, the pattern of local dependence in the conditional covariance among all item pairs is incorporated into the proximity measure. The new dimensionality sensitive proximity measure is computed as follows:

$$p_{ccor} = \sqrt{2(1-\frac{1}{\sum n_k} * \sum_{k=0}^{N-2} n_k * Corr_k)} \qquad (1)$$

where n_k is the sample size conditioned on the ability k, and $Corr_k$ is the correlation between the items pair conditioned on the ability k.

In the case of multidimensionality, the item pair covariance is conditioned on a unidimensional composite of multiple abilities (Roussos, Stout, & Marden, 1997). The covariance conditional on the composite abilities will be positive when the two items measure the same dimensions but negative when the two items measure different dimensions (Roussos, 1995; Roussos, Stout, & Marden, 1997). The patterns of positive and negative conditional covariances are the basis of new dimensionality sensitive proximity measures. However, since not all constructed-response items in a test have the same number of score categories, the conditional covariances may not be appropriate. The conditional correlation coefficients, which removes the variability of score differences by standardizing the item score range, were therefore used in place of the conditional covariances. The new dimensionality sensitive proximity measure, P_{ccor} is based on the conditional polychoric correlation between the two items. The $Corr_k$ is the polychoric correlation between two items conditioned on the remaining items. Since the observed variables were all ordinal, the use of ordinary product moment correlations based on raw scores was not recommended. Instead estimates of polychoric correlation were computed and used (Joreskog & Sorbom, 1988).

II. Cluster methods

There are various hierarchical cluster methods that are commonly used in cluster analysis: Single Link (SL) Complete Link (CL), Unweighted Pair-Group Method of Averages (UPGMA), Weighted Pair-Group Method of Averages (WPGMA), centroid method and Ward's mimimum variance Method (Ward's) . Previous research has provided inconclusive results regarding the "best" cluster method. Milligan (1981) found that no single method from the studies seemed to be more effective in term of recovery, although based on studies by Kuiper and Fisher (1975), Blashfield (1976), and Mojena (1977), Ward's method using Euclidean distance appeared to be the best algorithm for general data analysis (Millgan, 1981). However, other studies found that UPGMA provided superior recovery (Blashfield & Morey, 1980; Edelbrock, 1979;

Edelbrock & McLaughlin, 1980; Milligan & Isaac, 1980; Milligan, 1980). In some cases, distance-based measures were used, while in others, correlation-based measures were used.

In addition to the method that is used, Cronbach and Gleser (1953) stressed the importance of using appropriate proximity measures. While Milligan (1981) found that cluster methods were robust to the choice of proximity measures and concluded that the choice of cluster methods was more important than the choice of the proximity measure, Scheibler and Schneider (1985) concluded that the choice of clustering methods depended on the research purpose and the type of data in the study. As noted by Edelbrock and McLaughlin (1980), different cluster methods with different algorithms are designed to optimize different clustering criteria and thus give different clustering characteristics. Therefore, both the choice of proximity measures and cluster methods will depend on the type of data being analyzed.

III. Application of cluster analysis in testing

Hierarchical clustering analysis (HCA) has been used extensively in other areas like anthropology, psychology, engineering and market research. However, it has not been widely used in educational and psychological testing.

Miller and Hirsch (1992) used hierarchical cluster analysis to cluster IRT item discrimination parameters. The cluster analysis procedure was suggested as an alternative to conventional item factor analysis for investigating test dimensionality within a single test form and between alternate test forms. However, the IRT item parameters must be known before cluster analysis can be performed. Given that multidimensional calibration programs are necessary to perform these analyses and they are not readily available for polytomously scored data, the methodology of clustering item discrimination parameters is limited.

Hierarchical cluster analysis has been further explored and studied both theoretically and empirically in Roussos' dissertation and in his later research (Douglas, Kim, Roussos, Stout, & Zhang, 1995; Roussos, 1995; Roussos & Stout, 1995; Roussos, Stout, & Marden, 1997). The use of hierarchical cluster analysis in Roussos' simulation study identified highly correlated dimensions with as few as five items. As expected, he found that as the correlation between dimensions increased, the ability of the proximity measure to detect the dimensions underlying the test increased with test length. Although a promising tool for identifying groups of items reflecting different dimensions, Roussos concluded that the successful application of HCA is partly dependent on the formulation and/or choice of an appropriate proximity measure. Roussos' (1995; Roussos & Stout, 1995) research demonstrated that classical proximity measures are inappropriate for detecting dimensionally similar items because they tend to confound item difficulty with dimensionality. In his simulation study on dichotomous items, the new dimensionality sensitive proximity measures described by Roussos outperformed four classical proximity measures. Finally of the four selected cluster methods, only group average (UPGMA) performed well in the simulation with real dichotomous test response data.

IV. The stopping criterion in hierarchical cluster analysis

In cluster analysis, the selection of the number of clusters or partitions in the final solution is not always straightforward (Sneath & Sokal, 1973). The hierarchical cluster method does not indicate how many clusters underlie the data. In the hierarchical clustering process, a sequence of cluster solutions is obtained for each level. In order to obtain the optimal number of clusters, a stopping criterion or optimality criterion (Sneath and Sokal (1973)) is used. The cluster solution is usually evaluated by computing one or more available optimality criteria at each level. An optimal classification of objects occurs when the variances of between clusters to within clusters changes drastically between stages or levels of the cluster analysis.

Milligan and Cooper (1985) performed a comprehensive review of 30 criteria that are used to determine the optimal number of clusters in a data set and have found two criteria that performed best: pseudo F statistic by Calinski and Harabasz (1974) and the Je(2)/Je(1) ratio criterion by Duda and Hart (1973). The ratio index Je(2)/Je(1) is commonly seen as the inverse of pseudo t^2 statistics.

Although the Duda and Hart ratio criterion has some difficulty at the level of two clusters, the recovery performance in general has been good. The Calinski and Harabasz index, on the other hand, has performed rather consistently across the varying number of clusters. This statistic, however, performs well only if there are a few distinct spherical-shaped clusters (Jobson, 1992). Milligan and Cooper (1985) caution that the findings may be data dependent. However, they have felt that the ordering of the indices would not change too much even when a different data structure is used.

Summary

Assessing the dimensionality of test-response data is important for establishing the validity of test score inferences and the validity of using item response theory models. Hierarchical cluster analysis, in combination with a dimension sensitive proximity measure described by Roussos (1995), has been recommended as an alternative to factor-analytic methods for examining the dimensionality of a test. While a good deal of research has been presented on the use of cluster analysis, there is still a lack of consensus as to which methods are most effective. Scheibler and Schneider (1985) concluded that the choice of the cluster methods depends on the research purpose and the type of data being analyzed. An additional problem is that little research has been conducted in the context of test-response data.

The purpose of the present study was to examine the performance of cluster analysis in polytomously scored constructed-response tests. Such tests offer challenges to validating the underlying dimensionality because of the nature of the items and scoring criteria. The present study also evaluated cluster methods that were not studied by Roussos but have been found to be effective. Finally, a broader range in correlation among dimensions as well as unidimensional tests were considered. Unidimensional tests are relevant since many assessments are designed to be unidimensional so that IRT methods are applicable.

The specific research questions addressed in this study were:

(i) How well did the different cluster methods recover unidimensional and multidimensional data? Would different cluster methods over- or under-estimate the number of dimensions in unidimensional or multidimensional data?

(ii) How accurate were the two cluster stopping criteria?

(iii) How did the different cluster methods compare to the traditional exploratory factor analytic method?

(iv) Did test length affect the performance of different cluster methods?

(v) Did the correlation between dimensions affect the performance of different cluster methods?

Methodology

A Monte Carlo study was used to address the above research questions and considered the following factors: sample size, number of replications, proximity measure, cluster methods, stopping criteria, test length, number of dimensions and correlation between dimensions, and item and ability parameters. In this study only cluster methods, stopping criteria, test length, number of dimensions, and correlation between dimensions were chosen for manipulation.

(a) Sample size

In order to make the simulation realistic and manageable, the sample size was fixed at 3000 for all conditions. A sample size of 3000 was chosen to minimize any impact of the sample size on the results.

(b) Number of replications

The number of replications refers to the number of datasets that are generated for each combination of conditions under study. While much of the IRT-based MC research in the past typically used a single replication (e.g. Baker, 1990; Harwell & Janosky, 1991; Hulin, Lissak & Drasgow, 1982; Kim & Nicewander, 1993; Mislevy & Stocking, 1989; Swaminathan & Gifford, 1986; Yen, 1987), Harwell and colleagues (1996) advocate the use of multiple replications in order to increase the reliability and generalizability of the results. In this study, the ability of methods to recover the pre-specified dimensionality of the simulated test response data was being evaluated, the number of replications was fixed at 100 across the combination of factors being manipulated.

(c) Proximity measure

Since the classical proximity measure confounds item difficulty with dimensionality, the proximity measure developed by Roussos (1995) was the only proximity measure incorporated in this study. The only modification was the use of the conditional polychoric correlation instead of the conditional Pearson correlation. Since there are many dichotomously scored constructed-response items in the dataset, the use of conditional Pearson correlation may underestimate the relationship between those items. Conditional polychoric correlation was therefore used in place of conditional Pearson correlation in the dimensionality sensitive proximity measure.

(d) Methods for assessing dimensionality

Four methods of assessing the dimensionality structure were manipulated: three hierarchical cluster methods and exploratory factor analysis. The three hierarchical cluster methods had been selected based on past research in other areas of study. They are group average (UPGMA), centroid (UPGMC) and Ward's minimum variance methods. Each of the three cluster methods utilizes a different algorithm for determining the proximity between two clusters.

Exploratory factor analysis using polychoric correlations, which is a method traditionally used to assess dimensionality, was also evaluated. The maximum likelihood (ML) method was used since it provides a significance test for the "number of factors", and it also provides more accurate parameter estimates than the principal axis method (Hatcher, 1996). Since data were generated under conditions in which dimensions are uncorrelated and correlated, a varimax and promax rotation for each respective condition was conducted to facilitate interpretation.

SAS was used to perform both the cluster and factor analysis on the generated datasets. SAS PROC CLUSTER was used to perform the cluster analysis using Roussos' dimensionality sensitive proximity measure on each of the generated datasets. SAS PROC FACTOR was used to perform the ML factor analysis on each dataset. The proportion of variance accounted for was used to determine the underlying dimensionality of the one- and three-dimensional test.

(e) Stopping criterion (stopping rules)

Stopping criteria are needed to identify the number of clusters or dimensions in the cluster methods. The pseudo F statistic by Calinski and Harabasz (1974) and the ratio index [Je(2)/Je(1)] by Duda and Hart (1973), have been shown to be the two best performing indicators for determining the number of clusters (Milligan and Cooper, 1985), and were therefore chosen as the stopping criteria for the hierarchical clustering methods in the simulation study.

(f) Test length

Two test lengths, 12 and 24, were used. A test length of 24 was based on the New Standards Mathematics Reference Examinations. The average total test length of the three combined sub-tests is approximately 24 items for the various Mathematics Reference Examinations. Since other performance-based assessments such as NAEP have shorter tests with items as few as 12, a test length of 12 was chosen. Shorter (e.g. 6 items) or longer (e.g. 36 items) test lengths were not selected for manipulation since such test lengths are not typical of performance-based assessments.

(g) Number of dimensions and correlation between dimensions

The simulation study examined one- and three-dimensional data. Three-dimensional data was chosen to conform to constructed-response tests such as the New Standards Mathematics Reference Examination. This exam reports scores in three separate skill areas (Mathematics Skills, Mathematics Concepts, and Mathematics Problem Solving). Three-dimensional data was also chosen based on empirical results from estimating one, two, and three dimensional models for the exam using POLYFACT (Muraki, 1996).

The correlation of the three dimensions was also based on the relationship of the three sub-tests scores in the New Standards Mathematics Reference Examination. The values of the correlations that were used in this simulation were 0.00, 0.35 and 0.70. The correlation of 0.00 was used as the lower bound and baseline comparison. The correlation among the three factors in the real administration ranged from 0.69 to 0.83. A correlation of 0.70 approximated the relationships among New Standards sub-tests and was therefore chosen to represent a level in the manipulated factor. Finally a correlation of 0.35 was used to provide a comparison condition when the correlation between sub-tests was moderately low.

A one-dimensional condition was also evaluated in the present study. One-dimensional data were modeled to provide a comparison for the case that involves data appropriate for the application of unidimensional IRT models.

(h) Item and ability parameters

Using POLYFACT, item parameters for the simulation study were obtained by calibrating the middle school Form B Mathematics Reference Examination on 10,000 randomly selected examinees.

From the set of calibrated items (26 total items), 12 items were selected that represented a broad range in the types of items and the number of score levels for items on the New Standards Mathematics Reference Examination. The 12 items were selected based on a three-dimensional POLYFACT promax rotated solution. It should be noted that the item slope parameter estimates from POLYFACT correspond to an unrotated orthogonal solution. In order to obtain a pattern of estimates that conformed to approximate simple structure, the estimates for the uncorrelated and correlated three-dimensional condition slope parameters were transformed from the POLYFACT promax rotated solution (Muraki & Carlson, 1995). The promax rotated factor loading was used instead of the varimax rotated factor loading, because the promax rotated solution better approximated a simple structure factor solution. The 12 selected items were reanalyzed in POLYFACT to obtain the final item parameters used in the simulation study.

The 24-item test condition was obtained by replicating the set of 12 items. This served to reduce some of the random error when comparing the 12- with the 24-item model. Table 1 contains the set of 12 items and their associated parameters for the one-dimensional model. Table 2 contains the set of 12 items and their associated parameters for the three-dimensional model with uncorrelated and correlated dimensions. The highest absolute item discrimination is in bold to identify the dimension associated with the item.

Table 1
Slope and item-category threshold parameter estimates for the one-dimensional model.

		Slope			Item-category	
Item	Original Item no.	a1	b1	b2	b3	b4
Item 1	S202	0.54882	0.46909			
Item 2	S203	0.82198	0.53174			
Item 3	S204	0.67832	0.12981			
Item 4	S220	0.92390	0.77992			
Item 5	C211	0.75994	0.04136			
Item 6	C416	1.18408	-1.02871	-1.83245	-2.00558	
Item 7	C417	0.92260	-1.13282	-1.88452	-2.10966	
Item 8	C319	1.02345	-1.29256	-1.46657		
Item 9	P524	1.01483	2.49048	-0.46282	-1.16065	-1.73265
Item 10	P525	1.05424	2.19378	0.02274	-0.78911	-1.47746
Item 11	S313	0.72736	0.61184	0.45152		
Item 12	S314	0.91270	0.36579	0.17001		

<u>Note.</u> In the original item number, the first letter represents the skill areas (where S=Skills, C=Concepts, P=Problem Solving), the second numerical value represents the number of categories the item has, and the last two digits represent the item number.

Table 2
Slope and item-category threshold parameter estimates of the three-dimensional model for the uncorrelated and corrrelated dimensions

			Slope			Item-category		
Item	Original Item No.	a1	a2	a3	b1	b2	b3	b4
Item 1	S202	0.00870	0.01999	**0.58924**	0.48127			
Item 2	S203	-0.17499	-0.02783	**1.48390**	0.61136			
Item 3	S204	-0.14318	-0.02122	**1.13145**	0.13989			
Item 4	S220	0.23744	0.09685	**0.52685**	0.78015			
Item 5	C211	0.30112	-0.02500	**0.46931**	0.03851			
Item 6	C416	**1.59169**	0.11725	-0.20061	-1.12042	-1.99085	-2.17742	
Item 7	C417	**1.13316**	0.05144	-0.12775	-1.17924	-1.95681	-2.18861	
Item 8	C319	**1.01051**	-0.03324	0.06661	-1.32672	-1.50449		
Item 9	P524	**0.64897**	-0.02494	0.30716	2.51457	-0.47299	-1.18206	-1.76137
Item 10	P525	**0.90035**	-0.06427	0.21144	2.28296	0.01712	-0.83279	-1.55164
Item 11	S313	0.03951	**1.58271**	-0.09887	0.88976	0.65726		
Item 12	S314	0.08450	**0.94166**	0.18554	0.46293	0.21443		

<u>Note.</u> In the original item number, the first letter represents the skill areas (where S=Skills, C=Concepts, P=Problem Solving), the second numerical value represents the number of categories the item has, and the last two digits represent the item number.

For all conditions, ability parameters were randomly generated from a multivariate standard normal distribution with correlation among the dimensions equal to 0.00, 0.35, or 0.70.

Data generation

For each of the combination of conditions, RESGEN 2.1 (Muraki, 1992) was used to generate item response data using the defined item parameters and randomly generated abilities. One 1-dimensional item response dataset and three 3-dimensional item response datasets with a correlation of 0.00, 0.35, or 0.70 were generated. In order to check if the generated data conformed to the simulated one- or three-dimensional structure, the four sets of simulated item responses were checked using LISREL confirmatory factor analysis. The goodness of fit index (GFI) and adjusted goodness of fit index (AGFI) were consistent with the simulated dimensionality. These values ranged from 0.953 to 0.979 and 0.928 to 0.968 respectively. The RMSEA also validated the simulation procedure with values ranged from 0.0479 to 0.0733.

Given the item and ability parameters, response probabilities were first calculated using the multidimensional Graded Response Model (MGRM). The probabilities were then translated into discrete item responses by comparing the probabilities with random numbers drawn from a uniform distribution [0,1]. In the two-category response case, if the probability of a correct response for an examinee to an item is greater than or equal to the random number, a 1 is assigned to the items; otherwise a 0 is assigned. For greater than two category response items, if the random number falls between response categories k-1 and k, the item response of k is assigned to the item. The process was repeated with different random numbers for each item and for all examinees. The distribution of the item scores from the data generated using RESGEN was checked against the distribution of the item score in the real data through simple statistics (e.g., frequency distribution of each item). Consistency between the response distribution for the simulated and observed data was found.

In a simulation study, common item parameters and common seed values help minimize the effects of random error on parameter estimates. To reduce chance variations, common item parameters and common seed values may be used to generate a large data set so that replication can be randomly sampled from this population (Harwell, Stone, Hsu & Kirisci, 1996). However, this will create dependency among the datasets across the combinations of conditions. Alternatively, a seed can be used to generate a large set of item response data and subsets of non-overlapping blocks of examinees can be selected in a systematic order, creating independent replications. This was the approach taken in this present study. In the present study, a dataset of 300,000 examinees was generated and 100 data subsets of 3000 examinees defined the 100 replications for each set of condition were selected.

Determining the number of factors

In the case of hierarchical cluster methods the number of clusters was determined by computing the stopping criteria (pseudo F and t^2 statistics). For the pseudo F, a gradual monotonic increase in the pseudo F occurs as like items are joined and ends as dissimilar items or clusters are joined. The criterion value of cluster g immediately prior to this large decrease is the possible optimum cluster. In the case of pseudo t^2, a relatively large value of pseudo t^2 at cluster g would suggest that cluster (g+1), the relatively low criterion value, is the optimal cluster choice (Jobson, 1992).

In order to gauge the performance of EFA in recovering the underlying true dimensionality (1 or 3 dimensions), the proportion of variance accounted for by the initial eigenvalues was examined to determine the number of factors. A cut-point of 0.05 was used since this criterion has been found to identify dimensions of substantive importance (Hatcher, 1994). The eigenvalue greater-than-one criterion, or the Kaiser criterion, was also examined but not used to determine the number of factors. This criterion has been found to underestimate dimensionality (Cattell and Jasper, 1967; Browne, 1968) and it has been argued to be more appropriate for principal component analysis (Hatcher, 1994; Pedhazur & Schmelkin, 1991).

Dependent variables

Descriptive and inferential statistics were used to analyze the results of the simulation study. The dependent variable for this simulation study was based on the correspondence between the dimensionality identified through the analysis and the dimensionality under which the data were simulated. For each replication and within each combination of conditions, the match between the analysis and the simulated dimensionality was determined. It was coded 1 if the number of dimensions for the analysis matched the number of simulated dimensions; otherwise, it was coded 0. Therefore, the performance of the methods in recovering the underlying true dimensionality (1 or 3 dimensions) could be compared by using the number of times the method identified the true dimensionality across 100 replications.

Log-linear analysis was performed to assess the significance of the manipulated factors on the correspondence between the dimensionality uncovered in the analysis and the simulated dimensionality. Post-hoc analyses were also performed to determine which specific levels of a factor differed significantly.

Results

The purpose of this simulation study was to compare the recovery performance of three cluster methods (group average, centroid, Ward's), and the traditional factor analysis (EFA). In addition, several factors were manipulated to further evaluate the performance of the methods used to assess dimensionality: type of stopping criteria (pseudo F statistic, pseudo t^2 statistic) for cluster methods, length of test (12 items and 24 items), number of dimensions (1 and 3 dimensions), and magnitude of the correlation between dimensions (0.00, 0.35 and 0.70). One hundred datasets of 3000 item responses were simulated for each combination of test length, number of dimensions, and magnitude of correlation between the dimensions. The three cluster methods in combination with the two stopping criteria and EFA were then used to analyze the dimensionality of each simulated dataset.

Recovery of one-dimensional data – Descriptive results

Table 3 presents the percent agreement between recovered and simulated one-dimensional data for the method and test length variables. As can be seen in the one-dimensional data, the three cluster methods recovered the true dimensionality of the simulated data 23%, 98%, and 54% of the time respectively in the 12-item test. In the 24-item test, the three methods recovered the true dimensionality 12%, 82%, and 3% of the time. As test length increased the cluster methods overestimated the dimensions more frequently. Only the pseudo t^2 statistic stopping criterion was used since the pseudo F statistic can not be applied to the one-dimensional data. EFA, on the other hand, recovered the true dimensionality of the one-dimensional data in both 12-items and 24-items test 100% of the time.

Table 3
Percent agreement between recovered and simulated one-dimensional data for the method and test length variables.

		12 items		24 items	
Methods	Number of clusters	Pseudo t^2	Proportion variance > 0.05	Pseudo t^2	Proportion variance > 0.05
Average	1	23		12	
	≥ 2	77		88	
Centroid	1	98		82	
	≥ 2	2		18	
Ward's	1	54		3	
	≥ 2	46		97	
EFA	1		100		100

Recovery of three-dimensional data – Descriptive results

Table 4 demonstrates the correspondence between the number of dimensions recovered and the underlying true dimensionality (3-dimensions) across the 100 replications within each combination of conditions. In comparing the three hierarchical cluster methods, Ward's minimum variance method performed best when considering the three manipulated variables: type of cluster method, test length, and correlation between dimensions. The group average method was comparable to Ward's in terms of performance for correlated dimensions of 0.00 and 0.35. The centroid method performed markedly worst among the three cluster methods. With regards to the stopping criteria for the cluster methods, little difference was found between the use of the pseudo F versus pseudo t^2 statistic. As can be seen in the table, both stopping criteria were comparable in performance across the three cluster methods and across the three correlations. The one noteworthy exception was with Ward's method for the correlation of 0.70 in the 12-items test, where the pseudo t^2 statistic performed slightly better than pseudo F statistic. This result was consistent with Milligen's (1985) research, who found that both pseudo F and pseudo t^2 are the best performing stopping criteria and their rank ordering of the clusters do not change with different types of data.

Table 4
Percent agreement between recovered and simulated three-dimensional data for the method, correlation among dimensions, stopping criteria, and test length variables.

		12 items			24 items		
Correlation between dimension	Method	Pseudo F	Pseudo t^2	Proportion variance > 0.05	Pseudo F	Pseudo t^2	Proportion variance > 0.05
0.00	Average	100	100		100	100	
	Centroid	22	22		0	0	
	Ward's	100	100		100	100	
	EFA			100			100
0.35	Average	100	100		100	100	
	Centroid	4	0		0	0	
	Ward's	99	100		100	100	
	EFA			100			100
0.70	Average	80	79		82	82	
	Centroid	1	1		0	0	
	Ward's	92	97		100	100	
	EFA			89			100

When comparing the cluster methods to EFA, EFA performed nearly as well as Ward's method in recovering the true dimensionality of the multidimensional item response data. The one exception was for the correlation of 0.70 and the12-item test, where Ward's method using the pseudo F and pseudo t^2 identified the true dimensionality better than EFA. While the dimension recovery rate among the three cluster methods was similar for the correlation conditions of 0.00 and 0.35, the recovery rate differed as the correlation increased to 0.70. As the correlation increased to 0.70, the cluster methods less accurately recovered the true simulated dimensionality. As the correlation between dimensions increased the dimensions may seem more similar and behave more like unidimensional data.

As found by Roussos (1995), the recovery performance of the methods increased as number of items increased. However, this finding was limited to the case of correlated dimensions of 0.70. The one exception was the case of the centroid method, although it is of little interest given the poor performance of the method. The recovery rate of the 12-item test was very high among the two cluster methods (group average and Ward's) and EFA. Among these three methods, the recovery rate in the 12-item test ranged from 79% to 100%; while the recovery rate in the 24-item test ranged from 82% to 100%.

Table 5 presents the descriptive statistics for the eigenvalues and the proportion of variance accounted for in the EFA analyses of the 3-dimensional datasets. Only the statistics for the first four factors are reported. Although only the proportion of variance accounted for was used as the criterion for identifying the number of factors in EFA, both eigenvalues and proportion of variance accounted for were used to check the performance of the first three factors across the replications.

In Table 5, on average, the eigenvalues of the first three factors for the three-dimensional data were large compared to the fourth factor. The mean eigenvalues for the first three factors were also more prominent in the 24-item test when compared to the 12-item test. However as the correlations increased, the eigenvalues of the first factor for both 12-item and 24-item tests increased while the second and third factors decreased. The same was observed for the proportion of variance accounted statistics. A reasonable explanation for this was that the dimensions were becoming less distinct as the correlation increased. Finally, the average proportion of variance accounted for by the first three factors was greater than 0.05, which implied that, on average, the proportion of variance accounted for indicated the presence of three factors.

The results also demonstrated that the eigenvalue greater-than-one criterion would not have yielded the same dimensionality decisions as the proportion of variance accounted for criteria. As the correlation between dimensions increased to 0.70, the mean for the third eigenvalue for the 12-item test fell below 1.00. If the eigenvalue greater-than-one criterion was used to identify the number of dimensions in the test, the number of true dimensions would have been underestimated in most replications.

Table 5
Means and standard deviations of the eigenvalues and variances accounted for in the EFA across the replications.

		12 items				24 items			
		Eigenvalue		Proportion of variance accounted		Eigenvalue		Proportion of variance accounted	
Corr.	Factor	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.00	1	4.418	0.236	0.593	0.021	10.698	0.407	0.489	0.014
	2	3.100	0.170	0.417	0.019	7.846	0.354	0.359	0.013
	3	1.432	0.124	0.193	0.016	4.933	0.274	0.225	0.012
	4	0.039	0.029	0.005	0.004	0.251	0.036	0.011	0.002
0.35	1	6.044	0.253	0.768	0.015	14.495	0.438	0.650	0.011
	2	2.233	0.125	0.284	0.013	5.819	0.267	0.261	0.010
	3	1.047	0.093	0.133	0.012	3.539	0.199	0.159	0.008
	4	0.038	0.025	0.005	0.003	0.252	0.032	0.011	0.001
0.70	1	8.623	0.318	0.972	0.010	20.104	0.619	0.865	0.008
	2	1.054	0.085	0.119	0.009	2.828	0.157	0.122	0.007
	3	0.469	0.063	0.053	0.007	1.688	0.098	0.073	0.004
	4	0.050	0.027	0.006	0.003	0.256	0.033	0.011	0.001

Inferential Analyses

Recall that the dependent variable was dichotomized where a 1 was coded to indicate a match between the simulated dimensionality and dimensionality uncovered by the statistical method, otherwise a 0 was coded. Therefore, log-linear analyses were performed using PROC CATMOD in SAS to study the significance of the four manipulated variables: cluster methods, stopping criteria, test length, and correlation between dimensions.

Four log-linear models were estimated: main effects only, main effects plus all two-way interactions, main effects plus all two-way and three-way interactions, and the saturated model (main effects plus all two-, three-, and four-way interactions). Likelihood ratio statistics for an estimated model were used to evaluate the goodness of fit. Based on the comparison of goodness-of-fit statistics for the four models, the model with main effects plus all two-way interaction effects was the preferred model.

The analysis of variance table for the effects estimated in this particular model were used to identify the specific effects that are significant. These are presented in Table 6. As can be seen, the item, correlation and method main effects and correlation by method interaction effects were significant. The significant interaction effect showed that there was mutual dependence between the two variables: correlation and method. The item, method and correlation main effects were significant which implied that one level of method or correlation was significantly different from the next level. Based on Table 4, the centroid method may be the reason the method main effect was significant since the centroid method was very different from the other two cluster methods. The same could be observed for the correlation effect. The results for correlated dimensions of 0.70 were different from the other two levels (0.00 and 0.35) and thus may have accounted for the significant correlation main effect.

Table 6
Likelihood ratio tests for the effects estimated.

	Likelihood Ratio		
	Df	χ^2	P
Item	1	10740.62	0.0000
Corr	2	37.84	0.0000
Method	2	1108.51	0.0000
Stop	1	0.28	0.5994
Item*corr	2	0.23	0.8910
Item*method	1	.	.
Item*stop	1	0.01	0.9154
Corr*method	4	39.19	0.0000
Corr*stop	2	0.04	0.9819
Method*stop	2	0.36	0.8336

Since there were three different methods and three different correlations represented in the design, CONTRAST statements in SAS were used to compare the different levels in each of the two variables. Based on the results, the correlation condition of 0.00 (orthogonal level) was found to be significantly different from the other two levels (unorthogonal levels). In addition, the average and Ward's methods were similar while the centroid method was significantly different from the other two. However, these results should not be surprising given the results in Table 4. In the table the centroid method performed very poorly.

Since the centroid method was markedly different from the other two cluster methods the log-linear analyses were conducted disregarding the data for the centroid method. Upon rerunning the analyses without the centroid method, the contrast of correlation 0.00 versus 0.35 was not significant, but the contrasts of 0.00 versus 0.70, and 0.35 versus 0.70 were significant. As before, no differences were observed between the group average and Ward's method.

It can also be noted in Table 4 that most of the differences occurred for the correlation condition of 0.70. Therefore further analyses focused on this level to investigate the differences of the methods across the two test lengths. The log-linear analyses focusing on the correlation condition of 0.70 for the different methods (group average, Ward's, EFA) across the two test lengths (12 items, 24 items). Based on the results none of the methods was significantly different (average, Ward's, and EFA). Although it might seem that there was a difference in recovery rate across the two test lengths at the correlation of 0.70 in Table 4, the contrast statistics showed that the difference in test length was not significant for both stopping criteria (pseudo F and pseudo t^2). It should also be noted that the difference between the stopping criteria (number of match in pseudo F=97 and pseudo t^2=92) given a 12 item test and correlation condition of 0.70 was not significant (p=0.8302).

Further analyses

In view of the high recovery performance of EFA, further analyses were performed to explore conditions that might affect the performance of EFA in comparison with Ward's method. Ward's method was the focus of these analyses since it performed best across the three cluster methods. In order to gauge the effect of sample size, the sample size of each replication was reduced from 3,000 to 1,000. Also the effect of non-normal data was explored by generating multivariate gamma (positively skewed) data using RESGEN. However, since RESGEN cannot generate correlated skewed data, only uncorrelated skewed data were generated for comparison with uncorrelated standard normal data.

The results indicated that both EFA and Ward's method still performed well in recovering the underlying dimensions of standard normal data for sample sizes of 1,000. However, only Ward's method performed relatively well in recovering the underlying dimensions of positively skewed data. The recovery rate for EFA was 64% and the recovery rate for Ward's method was 85% and 82%, using the F and pseudo t^2 stopping criteria respectively.

All of the previous results examined the degree to which the true number of underlying dimensions was recovered. However, the degree to which the underlying factor structure was recovered, that is, the degree to which items were classified in their true underlying dimensions is also of interest. Recall that the true underlying factor structure of the 3 dimensional datasets involved 3 items that loaded exclusively on factor 1, 4 items that loaded exclusively on factor 2, 2 items that loaded exclusively on factor 3, and 3 items that loaded on factors 1 and 2. Thus, additional analyses were conducted in order to explore the item-level recovery of EFA versus Ward's method.

For the item level analysis, both methods, EFA and Ward's method for 3 dimensional datasets, did not perform well in recovering the true underlying factor structure. Cluster analysis using Ward's method was better than EFA in recovering the factor structure in standard normal data (42% in EFA, 74% for Ward's with F and t^2 stopping criteria). However, both EFA and Ward's methods did not recover the items for the respective dimensions well in positively skewed data (EFA=0%, Ward's with F=40% and Ward's with t^2=39%).

In examining the item level results, it was noted that three items were consistently misclassified and these same items also "loaded on" more than one dimension (items 4, 5, and 9 in Table 2). In order to examine the effect of these three items, these three items were removed. Once the three items were removed, the performance of the methods improved substantially. Full recovery in both Ward's and EFA on standard normal data. However EFA still did not recover the dimensions as well as Ward's method in the positively skewed data (EFA=5%, Ward's with F=99% and Ward's with t^2=98%). Ward's method may therefore be preferred with skewed data that has relatively simple structure.

Conclusion

The purpose of this study was to investigate hierarchical cluster methods as a potential dimensionality assessment tool for constructed-response tests. The specific research questions addressed in this study and the solutions to each question are presented below:

(i) How well did the different cluster methods (group average, centroid, Ward's cluster method) recover unidimensional and multidimensional data? Did different cluster methods over- or under-estimate the number of dimensions in unidimensional or multidimensional data?

Based on the simulation, only the centroid cluster method recovered the true dimensionality of simulated unidimensional data reasonably well and only in shorter tests (12 items). For all other conditions, the three cluster methods consistently overestimated the true dimensionality of the simulated data. One explanation to the overestimation may be that item parameters that were used to generate the one-dimensional data were not obtained from real one-dimensional data.

For 3 dimensional data, Ward's cluster method was the best performing method. Only the group average and Ward's cluster method recovered the multidimensional data well. Both group average and

Ward's cluster method were comparable in performance but Ward's cluster method performed better than group average when the dimensions were highly correlated. Among the three cluster methods, centroid method always overestimated the number of underlying dimensions, while the group average and Ward's method always underestimated the number of underlying dimensions in the three-dimensional case.

(ii) How accurate were the two cluster stopping criteria?

The two cluster stopping criteria (pseudo F and pseudo t^2 statistic) yielded comparable performance. However, pseudo t^2 statistic performed slightly better then pseudo F statistic for correlated dimensions of 0.70 using Ward's method in the 12-item test.

(iii) How did the different cluster methods compare to the traditional exploratory factor analytic method?

The traditional EFA method performed better than the cluster methods in unidimensional data and it performed as well as Ward's cluster method in recovering three-dimensional data. The one exception was for highly correlated dimensions and short tests, where Ward's method was superior. In addition, EFA did not recover the 3-dimensional data as well as Ward's cluster method in positively skewed data. Furthermore, Ward's performed better than EFA at item level recovery for both normal and positively skewed data.

(iv) Did test length affect the performance of different cluster methods?

Except for centroid cluster method, group average and Ward's method performed well for both 12 and 24 item tests. However, for highly correlated dimension (0.70), the group average and Ward's cluster method performed better in the longer tests.

(v) Did the correlation between dimensions affect the performance of different cluster methods?

As correlations between the dimensions increased, the dimensionality recovery rate decreased. The dimensionality was recovered best when the correlations were 0.00 and 0.35. As the correlation increased to 0.70, the cluster methods less accurately recovered the dimensionality.

Implications for practioners

The results of this study have implications for practioners who investigate the dimensionality of test response data. The results of this study indicate that EFA and Ward's method are recommended for long tests with moderately correlated (correlation = 0.35) simple structure normal data. Since EFA is more accessible than Ward's method with Roussos' proximity measure, EFA may be preferred. For tests that are shorter and which consist of dimensions that are more highly correlated, Ward's method may be better at uncovering the dimensionality than EFA. Also Ward's cluster method with either stopping criterion (pseudo F or pseudo t^2) is recommended for short tests with simple structure, and normally and positively skewed examinee score populations. However, despite the performance of the methods in uncovering the number of dimensions, there is some evidence that neither method is effective at uncovering the true interrelationships between items for tests with composite structure. Given that many tests exhibit such structure, this finding is disappointing and requires further research.

Limitations

There were a number of limitations in the simulation study. First, the use of polychoric correlations in the proximity measure might be problematic. As mentioned previously, in hierarchical cluster analysis, the user needs to identify an appropriate proximity measure. An appropriate proximity measure for use in assessing dimensionality has been developed by Roussos (1995). In Roussos' proximity measure for dichotomous items, Pearson correlations are used to calculate the conditional correlations between the items. Since this study included polytomous items with an ordinal score scale, conditional polychoric correlations were used in place of conditional Pearson correlations in calculations of the proximity measure. However, polychoric correlations can be unstable when the extreme cells in the contingency table have sparse frequencies. This forced some of the conditional polychoric correlation at the extreme scores to be unestimable. Whenever there were unestimable values in SAS, a missing value was excluded from the proximity measure calculation. This may have led to imprecision in the proximity measure.

Second, there were unestimable polychoric correlations in the correlation matrix for EFA. When there were missing values in the polychoric correlation matrix, the PROC FACTOR procedure stopped processing and an error message appeared in the log file. A value of 0 was used in place of the unestimable polychoric correlation. Although this occurred very infrequently, the problem of unestimable polychoric correlation therefore might also have added to imprecision in the EFA methods.

Third, the application of the stopping criteria is subjective. As the criterion values change very little for some cluster methods, identifying the optimum stopping criterion is difficult. This occurred primary for the centroid method. It would therefore be useful to establish the reliability of applying the stopping rules. This could be done by having a second person identify the optimum values and the consistency of the identification between raters recorded. Furthermore, of the two stopping criteria, only pseudo t^2 statistic can apply to both unidimensional and multidimensional data.

Finally, in any simulation study, the procedure does not fully capture the reality or all conditions in the testing situation. For example, nonresponse or other cognitive and behavioral processes were not modeled in the stochastic process. Therefore the absence of modeling all potential sources of error limits the generality of the Monte Carlo method and thus the generalizability of the results.

Further research

In hierarchical cluster analysis, the user must identify an appropriate proximity measure. Many classical proximity measures could not be used for ordinal data, such as item response data. In order to identify the dimensionality of the test, a dimensionality sensitive proximity measure with appropriate conditional covariance structure is needed. In this simulation study only conditional polychoric correlations were used in the proximity measure. The alternative conditional Pearson correlation that was not studied in this simulation could also be explored.

Tests with mixed item types consisting of both multiple-choice and constructed response have been gaining popularity. In this simulation, although various score levels (2, 3, 4, 5 level) of item responses were generated, they were all based on IRT models that do not estimate a guessing parameter. It may be of interest to see how the different dimensionality assessment methods perform with tests that involve mixtures of multiple-choice and constructed-response items.

Due to practical constraints, the simulation only examined a limited number of factors and levels of the factors. The test length factor could be further varied. In this simulation only two test lengths, 12-items and 24-items, were studied. Ward's method and EFA performed comparably for both test lengths. However, it is may be of interest to examine the performance of the methods with shorter tests (e.g., 6 items). Also, if the test length was varied for different combinations of multiple-choice and constructed-response items, generalizability of the results could be obtained.

Finally, other correlations between dimensions could be examined. In this simulation, only three levels of the correlation were studied. It may be of interest to study a broader range of more highly correlated dimensions (0.50-0.90).

Multidimensional scaling (usually abbreviated MDS) could also be used to study the underlying dimensions of test and inter-relationship between constructed-response items. MDS uses the proximity between objects to produce a spatial representation of the objects. Like cluster analysis, multidimensional scaling is an exploratory data analysis technique. Cluster analysis seeks to classify objects into groups using similarity measures derived from observed measurements. Multidimensional scaling seeks to determine the underlying dimensions that contribute to the perceived differences among the objects (Jobson, 1992). It could therefore be useful to compare the performance of MDS procedure with cluster analysis and EFA.

References

Baker, F.B. (1974). Stability of two hierarchical grouping techniques Case I: Sensitivity to data errors. *Journal of American Statistical association, 69*, 440-445.

Bennett, R.E. (1993). On the meanings of constructed response. In R.E. Bennett, & W.C. Ward, (Eds.) *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-28). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Blashfield, R.K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin, 83*, 377-388.

Blashfield, R.K., & Morey, L.C. (1980). A comparison of four clustering methods using MMPI Monte Carlo data. *Applied Psychological Measurement, 4*, 57-64.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1985). *Full-information item factor analysis* (MRC Rep. No. 85-1). Chicago: National Opinion Research Center.

Browne, M.W. (1968). A comparison of factor analytic techniques. *Psychometrika, 33*, 267-334.

Calinski, T. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3, 1-27.

Cattell, R.B., & Jaspers, J. (1967). A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. *Multivariate Behavioral Research Monographs*, 67-3.

Cronbach, L.J., & Gleser, G.C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456-473.

Douglas, J., Kim., H.-R., Roussos, L., Stout, W., & Zhang, J. (1995). *LSAT Dimensionality analysis for the December 1991, June 1992, and October 1992 administrations.* Unpublished manuscript, LSAC.

Duda, R.O. & Hart, P.E. (1973). Pattern classification and scene analysis. New York: John Wiley & Sons, Inc.

Edelbrock, C. (1979). Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. *Multivariate Behavioral Research, 14*, 267-384.

Edelbrock, C., & McLaughlin, B. (1980). Hierarchical cluster analysis using intraclass correlations: A mixture model study. *Multivariate Behavioral Research, 15*, 299-318.

Hambleton, R.K., & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

Harwell, M. R,, Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291.

Harwell, M., Stone, C.A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Hatcher, L (1994). A step-by-step approach to using the SAS system for factor analysis and structural equating modeling. Cary, N.C.: SAS Institute Inc.

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied PsychologicalMeasurement, 6*, 249-260.

Jobson, J. D. (1992). *Applied multivariate data analysis Volume II: Categorical and multivariate methods.* New York: Springer-Verlag.

Kim, J.K., & Nicewander, W.A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*, 587-599.

Kuiper, F.K., & Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics, 31*, 777-783.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.

Miller, T.R., & Hirsch, T.M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory. *Applied Measurement in Education, 5*, 193-211.

Milligan, G.W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45*, 325-342.

Milligan, G.W. (1981). A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research, 16*, 379-407.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*, 159-179.

Milligan, G. W., & Isaac, P.D. (1980). The validation of four ultrametric clustering algorithms. *Pattern Recognition, 12*, 41-50.

Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3-31.

Mislevy, R.J., & Stocking, M.L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal, 20*, 359-363.

Muraki, E. (1992). RESGEN version 2.1. Princeton, NJ: Educational Testing Service.

Muraki, E. (1996). POLYFACT. Princeton, NJ: Educational Testing Service.

Muraki, E., & Carlson J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73-90.

Muraki, E., & Engelhard, G. (1985). Full-information item factor analysis: applications of EAP scores. *Applied Psychological Measurement, 9*, 417-430.

Nandakumar, R., & Yu, F. (1996). Empirical validation of DIMTEST on nonnormal ability distributions. *Journal of Educational Measurement, 33*, 355-368.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Roussos, L. A. (1995). *A new dimensionality estimation tool for multiple-item tests and a new DIF analysis paradigm based on multidimensionality and construct validity.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Roussos, L. A., & Stout, W. F. (1995). *Effectiveness of using new proximity measures with hierarchical cluster analysis to detect dimensionality structure in simulated data.* Unpublished manuscript, University of Illinois at Urbana-Champaign.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1997). *Using new proximity measures with hierarchical cluster analysis to detect multidimensionality.* Unpublished manuscript, University of Illinois at Urbana-Champaign.

Roznowski, M., Tucker, L.R., & Humphreys, L.G. (1991). Three approaches to determining the dimensionality of binary items. *Applied Psychological Measurement, 15*, 109-127.

SAS/STAT User's guide, Version 6, Fourth edition Volume 1 (1993). Cary, NC: SAS Institute Inc.

Scheibler, D., & Schneider, W. (1985). Monte Carlo tests of the accuracy of cluster analysis algorithms: A comparison of hierarchical and nonhierarchical methods. *Multivariate Behavioral Research, 20*, 283-304.

Swaminathan, H., & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika. 51*, 589-601.

Yen, W. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.

AERA

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031536

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Assessing the dimensionality of constructed-response tests using hierarchical cluster analysis: A Monte Carlo study

Author(s): Brenda Siok-Hoon Tay-Lim, Clement A. Stone

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

↑

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

↑

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

↑

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: TaySiokHoon

Printed Name/Position/Title: Brenda Siok-Hoon Tay-Lim (Associate Research Scientist

Organization/Address: ETS, Rosedale Road MS 02-T, Princeton NJ.

Telephone: 609-734-1384

FAX: 609-734-5420

E-Mail Address: BLIM@ETS.ORG

Date: 5/31/00

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com